Incident Jul. 04, 2022: Performance degradation on VTEX IO infrastructure

Availability: **All accounts were available but faced increased errors regarding part of their requests**

% of clients affected: **All stores built with Store Framework**

Duration of incident: **1h51m**

Symptoms

From 19h22 UTC to 21h13 UTC, the VTEX IO platform faced performance issues that may have degraded some apps and caused a few timeouts and errors. Customers faced intermittent errors between 20h06 UTC and 20h57 UTC while navigating and ordering in stores that use Store Framework/VTEX IO.

Summary

At 19h12 UTC on July 4, we started rolling out traffic to a new version of the VTEX IO infrastructure that included some crucial improvements to the platform performance. This is usually a progressive maneuver; as we point traffic to the new infrastructure, we can monitor if there are any impacts through error and performance metrics.

At around 19h30 UTC, our alert system indicated an increase in response latency and error rate metrics related to some internal apps hosted in this new infrastructure. At this point, less than half of the traffic had been rolled out to the new infrastructure, and we started the rollback process. It was possible to see a slight impact on orders and sessions

starting from 19h22 UTC. At 20:01 UTC, we finished the rollback, and traffic pointed back to the old infrastructure version.

At 20h06 UTC, we identified an increase in the overall error rate on our platform again. After some investigation, we could see that the GraphQL Server was degraded. We started to apply some changes to mitigate it while spinning up new healthy infrastructure as a fallback plan if it didn't recover. Around 20h40, we could see that errors started to decrease as the applied changes took effect, and the system's health progressively recovered.

At 20h57 UTC, the error rate in our edge layer went back to normal such as the order and session behavior on the whole platform. Until 21h30, it was still possible to see a few errors for some apps, but we couldn't find evidence that it was still significantly affecting ordering and navigation.

Upon investigation, we discovered that a specific compute node of the GraphQL Server had connection problems. It downscaled in the time interval when the old infrastructure was not receiving production traffic. That increased overall latency and error rate due to timeouts trying to reach this service when traffic was redirected to it.

Important to note that while this incident partially impacted store environments, there was no impact on admin and dev environments related to that.

To be the trusted partner to your success, our team is working on follow-up actions to ensure that this incident does not happen again and that we identify and recover from future incidents faster. We are committed to improving our systems to guarantee a reliable and trusted platform.

Timeline

[2022-07-04 19:12 UTC] We started rolling out traffic to a new version of the VTEX IO infrastructure.
[2022-07-04 19:30 UTC] Our alert system indicated that the new environment was progressively degrading. We started the rollback process.
[2022-07-04 20:01 UTC] We finished rolling back traffic to the old infrastructure version.
[2022-07-04 20:06 UTC] Our alert system indicated an increase in response latency and error rate metrics related to the Store Framework rendering system.
[2022-07-04 20:25 UTC] We took some mitigation strategies to recover the system's health.

[2022-07-04 20:40 UTC] Errors started to decrease as the applied changes took effect and the system's health progressively recovered.
[2022-07-04 20:57 UTC] Errors related to the rendering services returned to expected levels.
[2022-07-04 21:00 UTC] We applied more changes in the infrastructure to mitigate a performance degradation that was still causing intermittent timeouts.
[2022-07-04 21:13 UTC] Error alarms were resolved.
[2022-07-01 21:27 UTC] We could confirm that all operations on the platform had been reestablished.

Mitigation strategy

We reestablished normal operations on VTEX IO infrastructure by moving traffic out of the degraded infrastructure.

Follow-up actions: preventing future failures

As a follow-up to this incident, we will work on improving our alarms to detect service connection degradation faster. We are reviewing the VTEX IO platform rollout/rollback tools and processes. We want to ensure that, during the following rollout of new infrastructure, the old environment will not downscale until everything is safely running in the new environment. This way, the recovery process will be faster, smoother, and safer whenever the rollout process fails.